AFRL-RH-WP-TR-2010-0097

Crosslingual Audio Information Retrieval Development

David M. Hoeferlin
Stephen A. Thorn

General Dynamics
Advanced Information Systems, Inc.
5200 Springfield Pike, Suite 200
Dayton OH 45431

April 2009

Interim Report for October 2005 to February 2009

Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Anticipate & Influence Behavior Division
Sensemaking & Organizational Effectiveness Branch
Wright-Patterson AFB OH 45433-7022

NOTICE AND SIGNATURE PAGE

AFRL-RH-WP-TR-2010-0097 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNED//
RAYMOND E. SLYH
Work Unit Manager
Sensemaking & Organizational
Effectiveness Branch

//SIGNED//
GLENN W. HARSHBERGER
Anticipate & Influence Behavior Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
April 2009	Interim	October 2005 – February 2009

4. TITLE AND SUBTITLE
Crosslingual Audio Information Retrieval Development

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
David M. Hoeferlin, Stephen A. Thorn

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
X0

5f. WORK UNIT NUMBER
7184X07C

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
General Dynamics
Advanced Information Systems, Inc.
5200 Springfield Pike, Suite 200
Dayton OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Anticipate & Influence Behavior Division
Sensemaking & Organizational Effectiveness Branch
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHXS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-WP-TR-2010-0097

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
88ABW cleared on 03 August 2010, 88ABW-2010-4152.

14. ABSTRACT
The Air Force Research Laboratory's Speech and Communication Research, Engineering, Analysis, and Modeling (SCREAM) Laboratory has a commercially-available system to encode, index, archive, and search multimedia events such as news broadcasts. The system is from a company that was formerly called Virage, which is now owned by a company called Autonomy. The Virage system contains a media encoder called a VideoLogger, and it has an audio indexing system from a company called BBN. The BBN audio indexing system gives the SCREAM Laboratory the capability to extract various metadata from audio and/or video content. This report discusses the development of a Virage Media Analysis Plug-ins (MAPs) to allow for translating text generated by the automatic speech recognition system (ASR) as well as a plug-in that allows other ASR or audio processing systems to be integrated with the Virage system.

15. SUBJECT TERMS Speech and Communication Research, Engineering, Analysis, and Modeling (SCREAM), Automatic Speech Recognition (ASR), Media Analysis Plug-in (MAP)

16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Raymond E. Slyh
U	U	U	SAR	18	19b. TELEPHONE NUMBER (include area code) NA

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

THIS PAGE LEFT INTENTIONALLY BLANK

TABLE OF CONTENTS

SUMMARY 1

1.0 INTRODUCTION 2

2.0 MAP DEVELOPMENT 3

2.1 Overview 3

2.2 SCREAM Virage Translator 3
2.2.1 MAP Component 3
2.2.2 Utterance Server Component 5
2.2.3 Machine Translation (MT) Component 6

2.3 SCREAM Virage Recognizer 7
2.3.1 MAP Component 7
2.3.1.1 Audio Resampling 7
2.3.1.2 Timing Problems 7
2.3.2 SONIC Server Component 8

3.0 MEDIA SEARCH INTERFACE 9

4.0 MULTILINGUAL CORPUS COLLECTION 11

5.0 RESULTS AND FUTURE WORK 12

5.1 Results 12
5.2 Future Work 12

REFERENCES 13

LIST OF ACRONYMS 14

LIST OF FIGURES

Figure **Page**

Figure 1: SCREAM Virage Translator Data Flow……………………………………...………... 3
Figure 2: MAP Configuration……………………………………………………………….4
Figure 3: VideoLogger with Utterance Track Displayed…………………………………………5
Figure 4: Utterance Server Output…………………………………………………………..6
Figure 5: SCREAM Virage Recognizer Data Flow………………………………………………7
Figure 6: SCREAM Media Search Web Interface………………………………………….10

SUMMARY

This report summarizes specific tasks completed by General Dynamics on the 711 Human Performance Wing/Anticipate & Influence Behavior Division, Sensemaking & Organizational Effectiveness Branch (711 HPW/RHXS) work unit 7184X07C, Crosslingual Audio Information Retrieval, for the period October 2005 to February 2009 under contract FA8650-04-C-6443.

1.0 INTRODUCTION

The Air Force Research Laboratory's Speech and Communication Research, Engineering, Analysis, and Modeling (SCREAM) Laboratory has a commercially-available system to encode, index, archive, and search multimedia events such as news broadcasts. The system is from a company that was formerly called Virage, which is now owned by a company called Autonomy. The Virage system contains a media encoder called a VideoLogger, and it has an audio indexing system from a company called BBN. The BBN audio indexing system gives the SCREAM Laboratory the capability to extract various metadata from audio and/or video content. The audio indexing system uses technologies such as automatic speech recognition (ASR), topic classification, speaker segmentation, speaker recognition, and named entity detection to extract information from audio. Specific information extracted includes spoken words, topic labels, identification of speakers, and entity tags such as person, location, organization, etc. The Virage system allows for the development of Media Analysis Plug-ins (MAPs), which can extend the media analysis capabilities of the VideoLogger.

This report discusses the development of a Virage MAP to allow for translating text generated by the ASR system as well as a plug-in that allows other ASR or audio processing systems to be integrated with the Virage system. Also discussed are the development of a search interface to allow for crosslingual audio information retrieval from foreign language media sources indexed by the Virage system as well as the collection of a corpus of foreign language materials to support the development of additional metadata detectors such as Interagency Language Roundtable (ILR) level, a United States Government-approved scale used to measure linguist proficiency level.[1]

An outline of this report is as follows. The next section describes the developed MAPs. Section 3 discusses the development of the search interface, while Section 4 describes the multilingual corpus collection. The final section summarizes the results and discusses future work.

[1] See `http://www.govtilr.org`

2.0 MAP DEVELOPMENT

2.1 Overview

While the Virage and BBN products provide useful capabilities, researchers with the SCREAM Laboratory desire to extend and enhance these capabilities as well as to create similar solutions for other languages not currently supported. Two capabilities were developed to interface with the Virage VideoLogger. The SCREAM Virage Translator sends the words from an ASR system to an external system for language translation, and the SCREAM Virage Recognizer sends audio to an external system for processing by ASR, speaker recognition, and/or other signal processing.

2.2 SCREAM Virage Translator

The SCREAM Virage Translator uses a VideoLogger MAP, an utterance server, and an external machine translation (MT) engine to translate the VideoLogger "Words" text track. As words become available from the audio indexing system in near real-time, the plug-in sends the words, identified speakers, and timing information to the utterance server. The utterance server groups words into sentence-like units, or utterances, based on the words, speakers, and timing information. Utterances are sent to an MT engine, and the translations are returned to the VideoLogger. Translation and utterance results are published to new VideoLogger text tracks called "Translation" and "Utterance." Figure 1 shows the data flow for the SCREAM Virage Translator system.

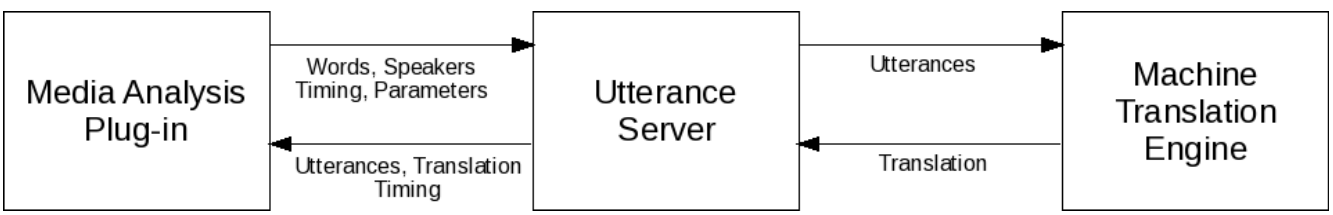

Figure 1: SCREAM Virage Translator Data Flow

2.2.1 MAP Component

The MAP component of the SCREAM Virage Translator is a Microsoft Windows Dynamic-Link Library (DLL) developed using the Virage VideoLogger Software Development Kit (SDK) and Microsoft Visual Studio C++ 6. The plug-in monitors the "Words" and "Speakers" text track in the VideoLogger and sends the words, speakers, and associated timing information to the utterance server via a Transmission Control Protocol /Internet Protocol (TCP/IP) socket connection. The plug-in receives utterances, translations and associated timing information from the utterance server and publishes the data in the VideoLogger interface.

The plug-in has several configuration parameters:

- Translation: This parameter identifies the translation to perform. Currently, this is limited to "Arabic to English" and "Chinese to English" based on the available BBN ASR capabilities integrated in the Virage system.

- Server: This parameter identifies the hostname of the utterance server.
- Port: This parameter identifies the TCP/IP port the utterance server listens on. The default value is 7890, but any valid TCP/IP port number is acceptable.
- Intraword Delay (ms): This parameter is used by the utterance server to divide the running word sequence into utterances. To calculate utterances, words are accumulated until the speaker changes or the delay between any two subsequent words exceeds the Intraword Delay parameter. A typical value might be 750, which is also the default value.

The configuration parameters are set in the VideoLogger using the “Media Analysis” tab of the Preferences window as shown in Figure 2. The Preferences window can be opened from the Options menu in the VideoLogger.

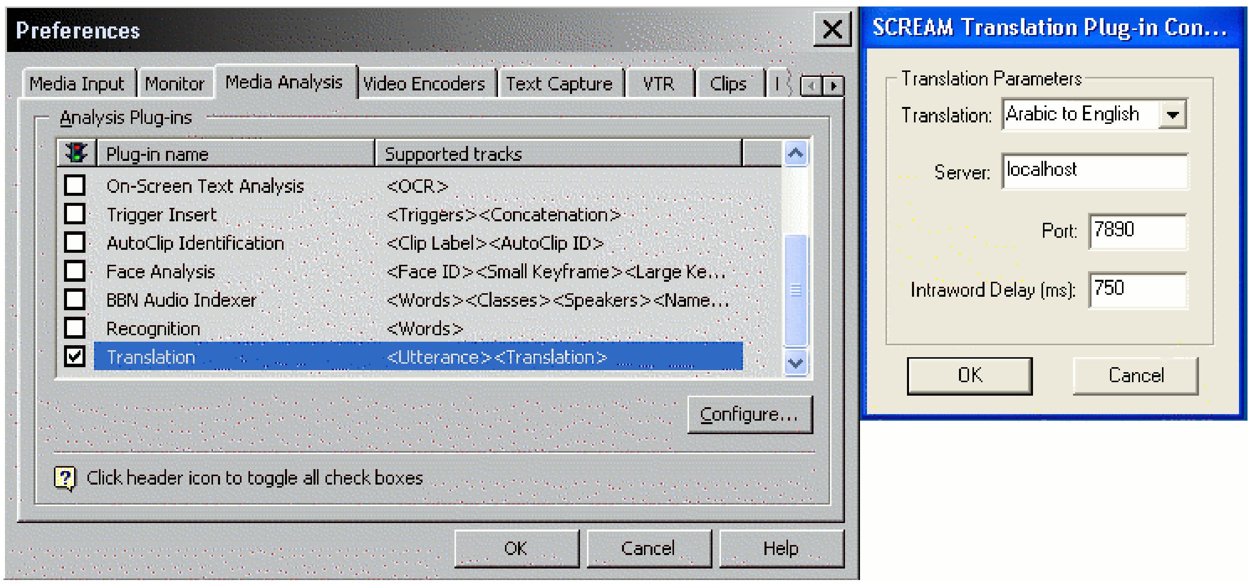

Figure 2: MAP Configuration

If the Translation plug-in is enabled, the VideoLogger will send content from the “Words” and “Speaker” tracks to the utterance server identified by the configuration parameters. The Translation (e.g. “Arabic to English”) and the Intraword Delay parameters are sent as well. The VideoLogger will receive utterances and translations from the utterance server and display them in the VideoLogger as shown in Figure 3.

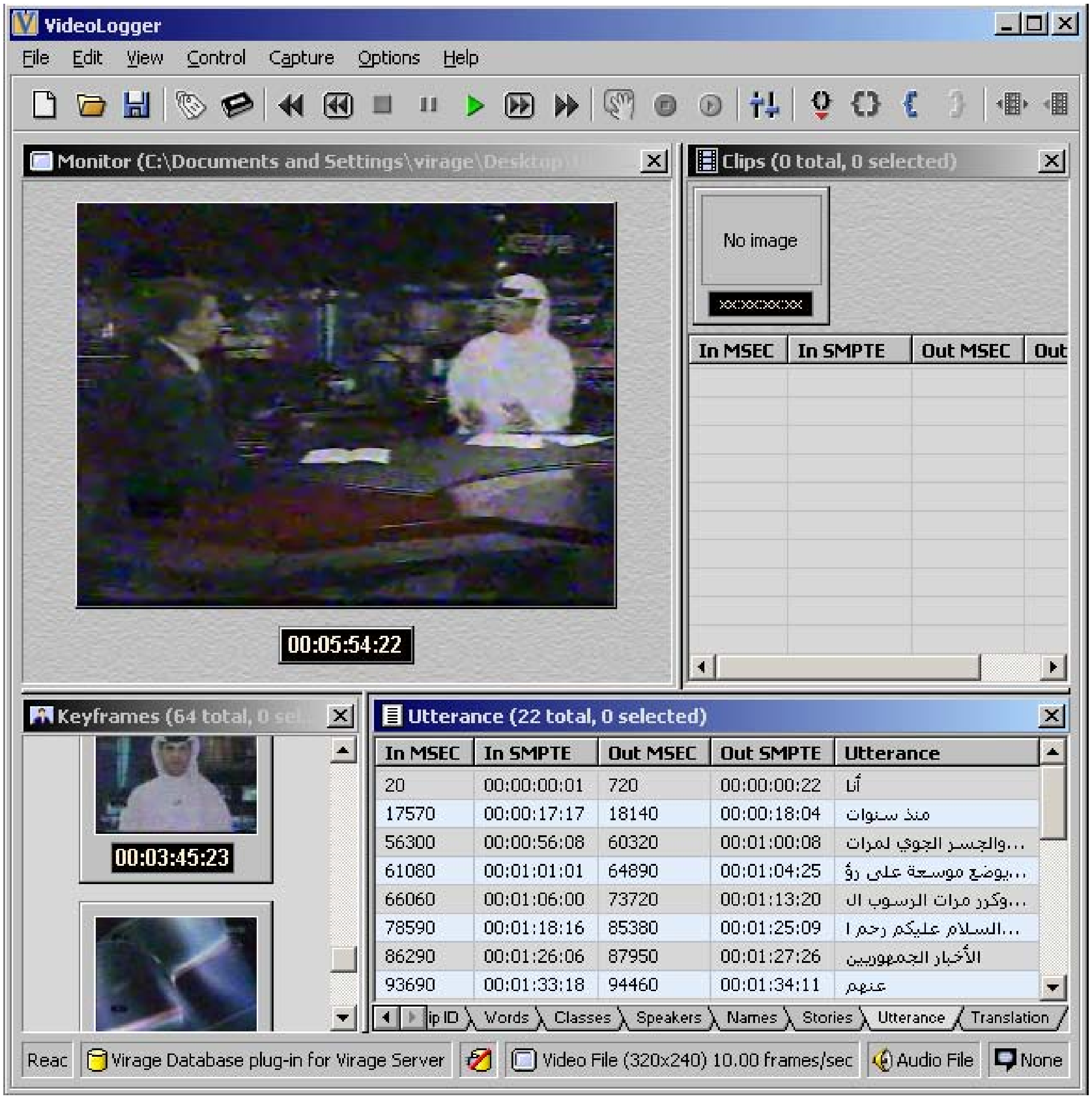

Figure 3: VideoLogger with Utterance Track Displayed

2.2.2 Utterance Server Component

The utterance server is typically run on the same host as the VideoLogger; however, it is implemented in the Perl programming language and can run on any host that has Perl installed. The utterance server receives "Words" and "Speakers" and associated timing from the VideoLogger. MT engines typically perform better when translating text with more context than they perform when just translating isolated words, so the utterance server is designed to collect words into sentence-like groups, or utterances. In order to determine the utterances, the server collects words that are from a particular speaker without long pauses. The length of a pause between words that will cause an utterance to end is the "Intraword Delay" as configured in the MAP component. Once a complete utterance is available, the utterance server connects to an MT engine to request a translation. The host providing the MT is configured near the top of the utterance server Perl script. The utterance server must connect to the MT engine on the

appropriate port for the desired translation language pair. These ports are configured in a file called `ports`, which should be in the same directory as the utterance server.

The `ports` file is a list of language pairs, ports and descriptions as shown below:

```
ar_en 10036 Arabic to English
zh_en 20444 Chinese to English
```

The language pairs and port numbers must match the appropriate ports used by the MT engines. Currently, the utterance server only supports the "SYSTRAN simple text-based TCP/IP protocol," so the language pairs and port numbers shown in the example `ports` file are some of those supported by SYSTRAN.

The utterance server displays some log information as data is received and translated. Depending on the capabilities of the console or window, the foreign language characters may not display correctly. However, this does not affect the results displayed in the VideoLogger. Example log output from the utterance server is shown in Figure 4.

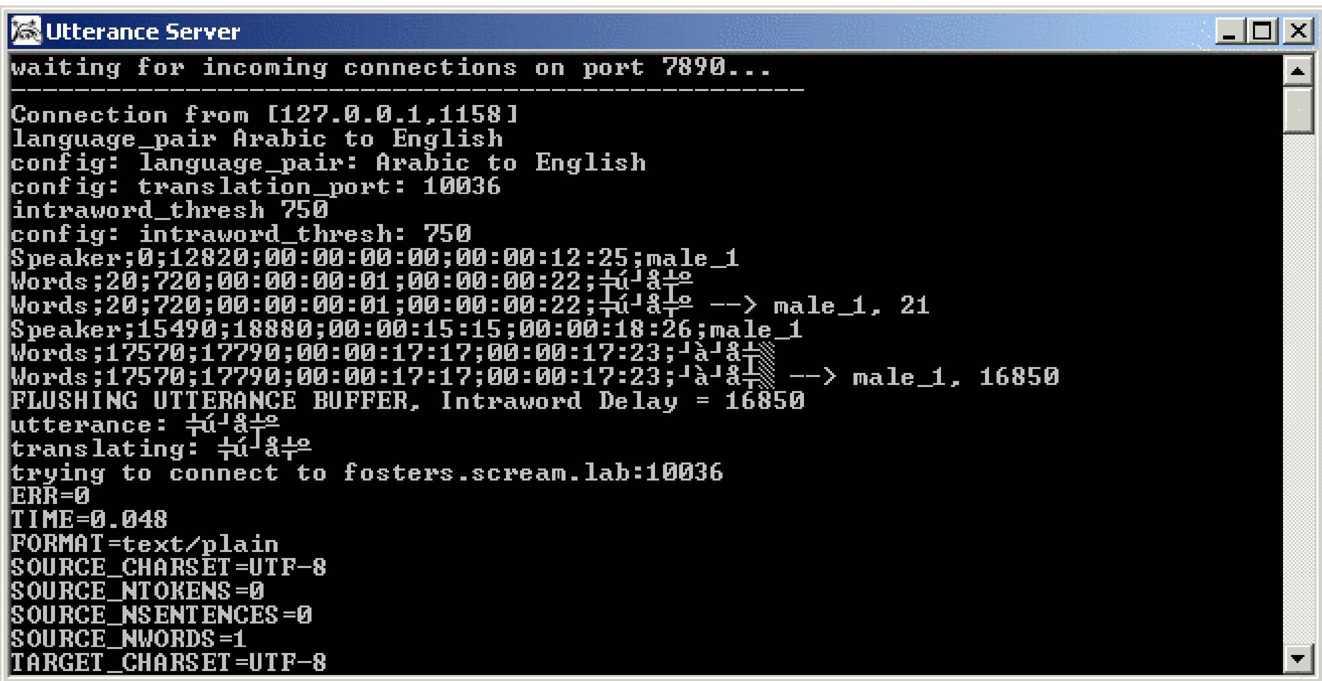

Figure 4: Utterance Server Output

After each utterance is sent to the MT engine, the utterance server waits for the resulting translation. Once the translation is received from the MT engine, the utterance server sends the translation and the corresponding utterance back to the VideoLogger where they are displayed under the appropriate tabs.

2.2.3 Machine Translation (MT) Component

The SCREAM Virage Translator currently uses SYSTRAN MT engines—specifically, the SYSTRAN Version USG 4.2 engines hosted on Solaris 8. If another MT engine were used, the utterance server would require modifications to interface with the desired engine.

2.3 SCREAM Virage Recognizer

The SCREAM Virage Recognizer uses a VideoLogger MAP to interface with an audio processing component, such as an ASR system like SONIC, a large vocabulary continuous speech recognition system developed at the University of Colorado at Boulder (Pellom 2001, Pellom & Hacioglu 2005). As the VideoLogger receives audio data from a multimedia event, the MAP sends a stream of audio data to an ASR server. After receiving enough data on which to perform ASR, the ASR server sends the recognized words back to the MAP. Figure 5 shows the data flow for the SCREAM Virage Recognizer system.

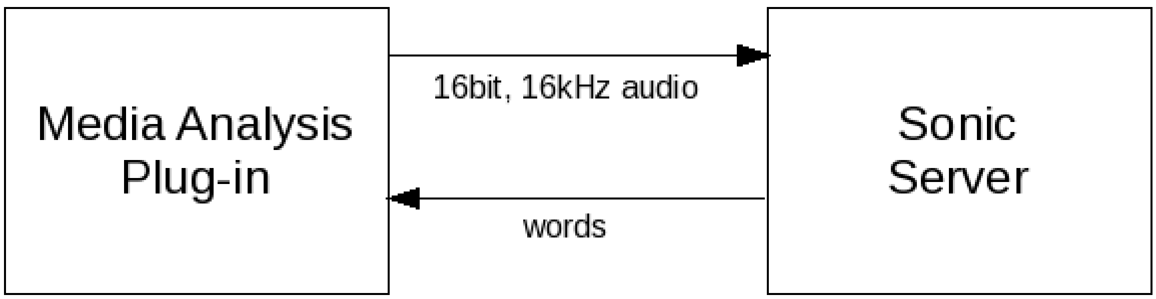

Figure 5: SCREAM Virage Recognizer Data Flow

2.3.1 MAP Component

The MAP component of the SCREAM Virage Recognizer is a Microsoft Windows DLL developed using the Virage VideoLogger SDK and Microsoft Visual Studio C++ 6. The plug-in requests the raw audio signal from the VideoLogger. Because the SONIC Server requires audio sampled at 16 kHz, the plug-in resamples the audio from the native VideoLogger rate, 22 kHz, to 16 kHz. The resampled audio is sent to the SONIC Server over a TCP/IP socket connection. Words recognized by the SONIC Server are received by the plug-in and written to the VideoLogger's media-analysis log file.

2.3.1.1 Audio Resampling

The MAP uses libresample, a real-time library for sampling rate conversion by Dominic Mazzoni.[2] Libresample is free software released under the Lesser General Public License (LGPL) from the Free Software Foundation. When raw audio becomes available to the VideoLogger, the MAP uses libresample to change the audio sampling rate as necessary to interface with the SONIC Server.

2.3.1.2 Timing Problems

While developing the SCREAM Virage Recognizer plug-in, we encountered errors when interfacing with the SONIC Server whereby the timing information for the individual recognized words was not correct. As a result, the current version of the Recognizer plug-in writes the recognized words to the VideoLogger media-analysis log file rather than to a text

[2] See http://ccrma.stanford.edu/~jos/resample/Free_Resampling_Software.html

track. If this timing issue is resolved in the future, then SONIC could be fully integrated. Other ASR servers can be fully integrated as long as they return the correct starting and ending times for each word.

2.3.2 SONIC Server Component

The SONIC Server component is the SONIC recognizer running in `live_mode` using the following configuration (stored in the `sonic.cfg` file):

```
-langmod_file           kb/wsj-5k-cnp.bin
-dictionary             kb/wsj-5k.lex
-phone_config           kb/phoneset.cfg
-acoustic_mod           kb/wsj-i.mod
-filler_file            kb/wsj.filler
-filler_penalty         0.0
-word_entry_beam        80.0
-state_beam             160.0
-word_end_beam          80.0
-lm_scale               25.0
-rescore_lm_scale       25.0
-word_trans_penalty     -12.5
-state_dur_scale        2.5
-short_word_penalty     0.0
-sample_rate            16000.0
-max_active_states      40000
-auto_end_point         1
-end_point_padding      125
-max_word_ends          400
-confidence             1
-confidence_am_scale    25.0
-live_mode              1
-push_to_talk           0
```

The SONIC Server is started using the following command:

```
SONIC/2.0-beta5/bin/i686-Linux/SONIC_server -g -port 5555 -c SONIC.cfg
```

The server can be tested by sending it a test audio file with the following command:

```
SONIC/2.0-beta5/bin/i686-Linux/SONIC_client -h localhost -p 5555 test.raw
```

3.0 MEDIA SEARCH INTERFACE

The rich metadata that results from the audio indexing performed by the Virage VideoLogger or similar systems can be useful for numerous applications. Crosslingual audio information retrieval to support language learning is one such application of interest to researchers in the SCREAM Laboratory. The media search interface application and multilingual corpus collection (discussed in the next section) were projects conducted to support the language learning application.

The SCREAM Media Search is a web based application to search the multilingual, multimedia data collected, encoded, and indexed with the Virage VideoLogger or similar system. The media search application demonstrated a method of searching the metadata for specific keywords in English or the foreign languages supported by ASR systems and displaying the results with additional analysis data such as vocabulary coverage ranking.

Figure 6: SCREAM Media Search Web Interface

The media search engine was developed using the full-text search capabilities of MySQL[3]—namely, Boolean search, natural language search, and query expansion search. Full-text searching is performed using "MATCH() ... AGAINST" syntax. "MATCH()" takes a comma-separated list that names the columns to be searched. "AGAINST" takes a string to search for and an optional modifier that indicates what type of search to perform. The search string must be a literal string, not a variable or a column name.

A Boolean search interprets the search string using the rules of a special query language. The string contains the words to search for. It can also contain operators that specify requirements such that a word must be present or absent in matching rows, or that it should be weighted higher or lower than usual.

[3] See `http://www.mysql.com`

A natural language search interprets the search string as a phrase in natural human language (i.e., a phrase that could occur in free text); there are no special operators. However, a stopword list (i.e., a list of common words such as “the,” “and,” “a,” and “an” that do not carry much information content for retrieval purposes) is applied, so that the presence or absence of the stopwords in the query or the database does not affect the search results. In addition, words that are present in 50% or more of the rows are considered common and do not match.

A query expansion search is a modification of a natural language search. The search string is used to perform a natural language search. Then, words from the most relevant rows returned by the search are added to the search string, and the search is performed again. The query returns the rows from the second search. A query expansion search can boost recall (i.e., the percentage of relevant documents that are returned) at a cost of lowering precision (i.e., the percentage of returned documents that are relevant).

The information in the database is searchable by keywords, but the search can be narrowed to search only particular tracks. Selectable tracks for searching include: Closed Caption, Names, Speakers, Speaker Identification (ID), Speech, Stories, Translation, Utterance, Words, and All Tracks.

The search results include links to the original video streams according to the time-code values stored within the database. The videos are available to play as a full clip of the event or as a user-defined portion of the clip according to the search parameters.

4.0 MULTILINGUAL CORPUS COLLECTION

The collection of a multilingual corpus was initiated for use in developing detectors for the ILR level as well as other metadata. The corpus was created by retrieving lessons from the Global Language Online Support System (GLOSS), [4] a web site provided by the Curriculum Development Division of the Defense Language Institute Foreign Language Center (DLIFLC). GLOSS language lessons are developed for students and Department of Defense linguists to support language learning and sustainment in reading and listening using authentic materials such as magazine articles, TV and radio broadcasts, and interviews.

At the time of the corpus collection, the GLOSS site provided materials in 27 languages, grouped by ILR proficiency level, skill modality, competence, and topic. The ILR proficiency level for each lesson was labeled by trained raters according to ILR standards. The ILR scale consists of six "base levels" ranging from 0, No Proficiency, to 5, Functionally Native Proficiency, with intervening "plus levels" that indicate when the required proficiency level substantially exceeds one base skill level, but does not fully require the criteria for the next "base level." The lessons retrieved from the GLOSS site consisted almost entirely of lessons rated in the 2, 2+, and 3 levels. The skill modality refers to whether the lesson is based on listening or reading. The competence refers to whether the lesson primarily focuses on lexical, discourse, structural, or socio-cultural content of the material. The topics covered in the lessons are: Culture, Economy, Environment, Geography, Military, Politics, Science, Security, Society, and Technology.

A list of available lessons was created for each language using the GLOSS naming schema, and the lists were used in a web scraping tool to collect the relevant files. Each lesson consisted of multiple Hyper Text Markup Language (HTML), image, and multimedia files. The collected HTML files were edited to pull out the source text and the English translations for further use. In total, the amount of captured information measured over two gigabytes with nearly 21,000 files.

[4] See `http://gloss.lingnet.org`

5.0 RESULTS AND FUTURE WORK

5.1 Results

The SCREAM Virage Translator successfully integrates the Virage VideoLogger, BBN audio indexing system, and SYSTRAN MT engines to provide language translations of live or recorded multimedia events. Although the system currently only handles Arabic to English and Chinese to English translations, it could be easily extended to additional languages if the necessary ASR and MT capabilities were available.

The combination of the MAP and the utterance server was a design to keep the MAP minimalistic and increase flexibility. This flexibility could be enhanced by a more general version of the MAP that would allow any Virage VideoLogger text track to be retrieved instead of just the "Words" track.

The SCREAM Virage Recognizer successfully demonstrates the use of the SONIC recognizer to perform ASR on Virage VideoLogger media events. Development was not 100% completed after the timing problems with the interface to the SONIC Server were discovered. The ASR results from the SONIC Server were written to the VideoLogger media analysis log file instead of being published as a text track in the VideoLogger interface as publishing a text track in the VideoLogger requires a correct starting and ending time for each element.

A search interface was developed that allowed for crosslingual audio information retrieval based on the metadata in the Virage database. The search can be narrowed to search only particular tracks.

A multilingual corpus was collected to facilitate the development of detectors for ILR level and other metadata. If these detectors are integrated into the Virage system to provide additional metadata tracks, then these tracks can be made available to the search interface.

5.2 Future Work

One potential method of solving the SCREAM Virage Recognizer timing problem while still using the SONIC Server would involve sending segments of audio to the SONIC server via individual TCP/IP socket connections. The length of each audio segment could be used to calculate the starting and ending time for the group of words recognized for each segment. This method would also require the MAP component to implement a robust speech/silence detector to avoid segmenting the audio during active speech. Using individual TCP/IP socket connections for each audio segment would also incur additional network overhead as a network socket would be opened and closed for each speech segment. While the SCREAM Virage Recognizer currently only communicates with the SONIC Server for ASR, it could be extended easily to support other ASR systems.

Future work by SCREAM Lab researchers will focus on developing various metadata detectors, such as detectors for ILR level. When these detectors are complete, they can be integrated into the Virage system with plug-ins and their associated metadata tracks can be provided to the search interface.

REFERENCES

Pellom, B. "*SONIC: The University of Colorado Continuous Speech Recognizer,"* Technical Report TR-CSLR-2001-01. University of Colorado, Boulder, CO, March 2001.

Pellom, B., & Hacioglu, K. "Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Hong Kong, April 2003.

LIST OF ACRONYMS

ASR	Automatic Speech Recognition
DLIFLC	Defense Language Institute Foreign Language Center
DLL	Dynamic-Link Library
GLOSS	Global Language Online Support System
HTML	Hyper Text Markup Language
ID	Identification
ILR	Interagency Language Roundtable
LGPL	Lesser General Public License
MAP	Media Analysis Plug-ins
MT	Machine Translation
SCREAM	Speech and Communication Research Engineering, Analysis and Modeling
SDK	Software Development Kit
TCP/IP	Transmission Control Protocol/Internet Protocol
711 HPW/RHXS	711 Human Performance Wing/Anticipate & Influence Behavior Division, Sensemaking & Organizational Effectiveness Branch